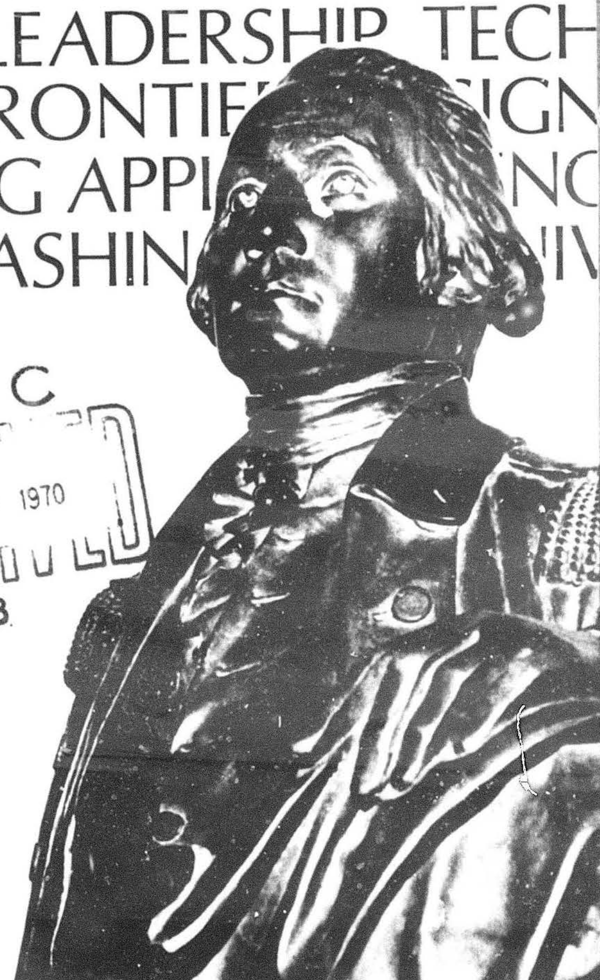
THE GEORGE WASHINGTON UNIVERSITY

STUDENTS FACULTY STUDY R
ESEARCH DEVELOPMENT FUT
URE CAREER CREATIVITY CO
MMUNITY LEADERSHIP TECH
NOLOGY FRONTIE SIGN
ENGINEERING APPL NC
GEORGE WASHIN IV



INSTITUTE FOR MANAGEMENT SCIENCE AND ENGINEERING

SCHOOL OF ENGINEERING AND APPLIED SCIENCE

A STATISTICAL ANALYSIS OF THE ENGINEERING APPROACH TO NAVY SHIPBUILDING COST ESTIMATION

by

K. C. Yu

Serial T-240
30 June 1970

The George Washington University
Institute for Management Science and Engineering
Program in Logistics

Contract N00014-67-A-0214
Task 0001, Project NR 347 020
Office of Naval Research

THE GEORGE WASHINGTON UNIVERSITY
School of Engineering and Applied Science
Institute for Management Science and Engineering

Program in Logistics

Abstract
of
Serial T-240
30 June 1970

A STATISTICAL ANALYSIS OF THE ENGINEERING APPROACH TO NAVY SHIPBUILDING COST ESTIMATION

by

K. C. Yu

In this study, the feasibility of developing regression models to predict the total cost of a Navy ship using the physical weights of the ship components as independent variables was investigated. The various forms of regression analyses fall under the following three categories:

(1) Linear multiple regression analysis.

(2) Non-linear multiple regression analysis.

(3) Adding-up process, which is an aggregation of two-variable regression analyses.

It was found that the linear model is preferable over the non-linear model and the adding-up process. If the samples are properly selected, linear models which are statistically significant can be derived. Given its superiority over the other two models, the degree of accuracy of the linear model is still not high enough to produce a dependable point estimation for the total cost of the ship.

A STATISTICAL ANALYSIS OF THE ENGINEERING APPROACH TO NAVY SHIPBUILDING COST ESTIMATION

By

K. C. Yu

Bachelor of Science
Mapua Institute of Technology, 1956

A Thesis Submitted to the School of Government and Business Administration of the George Washington University in Partial Fulfillment of the Requirements for the Degree of Master of Business Administration

June 1970

Thesis Directed By
John H. Norton, PH.D
Associate Professor of Business Administration

ACKNOWLEDGMENT

This research was sponsored by The George Washington University, Program in Logistics under contract to the Office of Naval Research. I wish to express my gratitude to Professor John Norton, my thesis director, for the great amount of time he provided in guiding the various stages of this research; to Professor Henry Solomon, Program in Logistics, for providing the framework for the direction of this research and his many valuable suggestions; and to Professor Robert Holland for his helpful suggestions. Appreciation is also expressed to Captain J. E. Hamilton for his assistance on matters related to Navy ship data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Subject, Objectives, Scope

Shipbuilding cost estimation has been a subject of great importance to the U. S. Navy. Past experience shows that it is difficult to estimate shipbuilding costs accurately. This is true because the aspects of ship cost are extremely complex. It encompasses numerous variables which are not fully identified and the relationships among these variables are not fully understood.

This study consists mainly of statistical analyses of Navy ship cost estimate data. The first objective of this study is to identify some of the important engineering variables and to investigate their significance and relationships with respect to cost estimates. The second objective is to test the effectiveness of some methodologies chosen to achieve the first objective.

The data base available to this study consists of cost estimate data from a collection of contractor's bid proposals for Navy ships which were to be built. Therefore this study focuses on the earlier cost estimates, i.e., during the conceptual design stage. These estimates are important to the Navy as they serve as major factors in determining whether or not the ship will be included in the shipbuilding program. Estimation of the alteration, repair, and maintenance costs will not be included in this study.

It should be emphasized that the data used in this study are not actual production cost data. They are estimates from contractors of what their production cost will be and what their profit will be to arrive at an estimated total cost. The justifications of using these data are first, they are the only data available and second, the results of the investigation will serve the practical purposes of evaluating on what basis the contractors prepare their bid proposals and predicting how much the contractors will bid. In this context, the contractors can be regarded as manufacturers and bidders. Their bidding characteristics will primarily be a function of the engineering and economic factors. Therefore the engineering and economic aspects of ship cost with emphasis on the former will be the center of interest in this study.

Theoretical Basis

The theoretical basis which serves as a guide line to this study and possible future extension of this study is derived from the work of Professor Henry Solomon.[1] Some of the important concepts of Professor Solomon are outlined below.

"The earlier cost estimate requirements present great difficulties for at least two reasons. (1) the relative scarcity of detailed technical information on which to base the estimates and (2) the many uncertainties concerning changes in technological, mobilization and economic conditions, since the path from the design stage to the completion of the construction stage requires a long time interval."

---

[1] Henry Solomon, "Estimating Cost of New Construction Navy Ships" Program in Logistics, George Washington University, 1969.

"Cost estimates may be used for two purposes. One is to serve as an input into program analysis (i.e., trade-off analysis) to serve as a guide for choosing among alternative designs, and the second is to determine an actual budgetary requirement. Ideally, the same cost estimate may serve both purposes, however, the demand for accuracy may in fact not be the same. It was often suggested that for program analysis, the absolute cost value need not be accurate provided that the relative costs are accurate among the alternative designs. In analyzing budgetary requirement, higher accuracy in the absolute cost is required."

"It is convenient to partition approaches to cost estimation into two types: engineering and economic. By no means are these approaches mutually exclusive, however, they do differ in emphasis. The engineering approach refers to the estimate of costs of design entities which form a ship or some multiple of the same ship without specific reference to the over-all shipbuilding program over time and/or the economic conditions of the industry during the relevant time period. The economic approach refers to studies which attempt to deal explicitly with the industry and general economic factors such as relative utilization of factors of production, scale of output, industrial structure, etc., without specific reference to individual product design. Obviously the need exists to integrate both approaches but this integration may not come about by simply using both. Each approach must likely be modified in form and content to effect desired integration leading to a total model. This is an ambitious objective and one which cannot be accomplished within the near future. A more modest goal is to effect improvements where necessary for each approach while recognizing and formulating a more complete model."

DATA

The data base for this study consists of two files. The cost file and the weight file.

THE COST FILE

The cost file is a collection of bid proposals prepared by various contractors between the year 1954 to 1966. There are 989 records in the file. Each record represents a bid proposal prepared by a particular contractor in response to a particular "invitation to bid" issued by the Navy for the construction of a new ship. Among the total 989 records there are 127 "invitations to bid". Each "invitation to bid" involves only one distinct ship prototype although several may be built.

Among the 989 records there are 29 different types of ships. Each type of ship may include one or more different classes (each class is identified with one distinct prototype). There are a total of 58 different classes.

Each record includes among other information the following costs of ship components which are important to this study.

Hull structure cost
Propulsion cost
Electric plant cost
Communication and control equipments cost
Auxiliary system cost
Outfit and furnishing cost
Armament cost
Profit
Total Cost

It should be noted that these costs are contractor's estimated costs.

The Weight File

The weight file consists of 143 records. There are 27 different types of ships included in the file. Each type of ship may include one

or more different classes. The different classes of ships under a particular ship type in the weight file unfortunately do not correspond exactly to those in the cost file. Each record contains for one particular class of ship the following physical weights of the ship components in tons.

Hull structure weight
Propulsion weight
Electric plant weight
Communication and control equipments weight
Auxiliary system weight
Outfit and furnishing weight
Armament weight

Methodology

The basic statistical tool to be used in this study is regression analysis. The fundamental approach is to test the feasibility of deriving regression equations which can accurately predict the total cost of the ship base on the engineering weight components of the ship.

It should be noted that in the traditional engineering approach of estimating ship cost, a ship is usually divided into seven weight groups. They are:

Hull structure
Propulsion
Electric plant
Communication and control equipments
Outfit and furnishings
Auxiliary systems
Armament

In this study, the seven weight groups or their various transformations are primary elements which are to be used as independent variables of the regression analyses.

There are two important aspects of regression analysis which this study must deal with.

(1) The selection of samples, i.e., the determination of sample size and the basis of stratification. The stratifications to be considered are:

(A) Stratification by the hull weight of the ship.

(B) Stratification by propulsion weight of the ship.

(C) Stratification by usage of the ship.

(2) The form of the regression model (detailed descriptions of the various forms will be presented in the next coming section). The different forms of the regression model are:

(A) The adding-up process[1] which is an aggregation of two-variable regression analyses.

(B) Linear multiple regression analysis.

(C) Non-linear multiple regression analysis which can be further subdivided into.

(a) Dependent variable in non-linear form and independent variables in linear form.

(b) Dependent variable in linear form and independent variables in non-linear form.

(c) Dependent variables in linear form and independent variables in both linear and non-linear form.

(d) Both dependent and independent variables in non-linear form.

The guide line for the direction of the experimentations in this study is based on the belief that a better regression equation can

---

[1]The "adding-up process" is the term adopted to describe the method. It is not standard statistical terminology.

be derived by proper selection of the above two factors, i.e., proper selection of samples and proper selection of regression forms. Therefore the effort is directed at determining which combination of these two factors will bear better results and how significant are these results.

Detailed Descriptions of the Different Forms of Regression Analysis

The Adding-up Process.

In the adding up process of estimating ship cost, two-variable regression equations are derived for each of the seven weight groups with component cost[1] as the dependent variable and component weight[2] as the independent variable.

The two-variable regression equation takes the form

$$c_{ij} = a_j + b_j W_{ij} + e_{ij} , \quad j = 1 \text{ to } 7 , \quad i = 1 \text{ to } n .$$

c_{ij} = known component cost of the seven weight groups.

a_j , b_j = regression parameters to be determined.

W_{ij} = component weights of the seven weight groups.

e_{ij} = the residual, n = number of observations.

---

[1] The term "component cost" refers to the cost of the weight groups, i.e., hull cost, propulsion cost etc.

[2] The term "component weight" refer to the physical weight of the weight groups, i.e., hull weight, propulsion weight etc.

After the a_j's and b_j's are determined by the regression analysis, which will now be called $\hat{a}_j$ and $\hat{b}_j$, the prediction equation for the component cost of the seven weight groups becomes

$$\hat{c}_{ij} = \hat{a}_j + \hat{b}_j W_{ij}, \qquad j = 1 \text{ to } 7 , i = 1 \text{ to } n$$

$\hat{c}_{ij}$ = the estimated component cost.

The total cost of the ship can then be estimated by using the equation

$$\hat{Y}_i = \sum_{j=1}^{7} \hat{c}_{ij} + DEC_i + CSC_i + profit_i , \quad i = 1 \text{ to } n$$

$\hat{Y}_i$ = estimated cost of the ship.

DEC_i = design and engineering cost.

CSC_i = construction cost.

Linear Multiple Regression Analysis

In the linear multiple regression analysis, the general form of the equation is as follows:

$$Y_i = b_o + \sum_{j=1}^{7} b_j W_{ij} + e_i$$

Y_i = the known total cost of the ship.

b_o = regression constant to be determined.

b_j = regression coefficients to be determined.

W_{ij} = component weights of the seven weight groups.

e_i = the residual.

Non-linear Multiple Regression Analysis

The equations for the non-linear multiple regression analysis may take the following forms:

$$(1) \quad Y_i = b_o + \sum_{j=1}^{7} b_j \, f(W_{ij})^1 + e_i$$

$$(2) \quad f(Y_i)^1 = b_o + \sum_{j=1}^{7} b_j \, W_{ij} + e_i$$

$$(3) \quad Y_i = b_o + \sum_{j=1}^{7} b_j \, W_{ij} + \sum_{k=1}^{m} b'_k \, X_{ik} + e_i$$

X_{ik} = any non-linear forms of the independent variables such as the squares or the cross products.

m = number of X_{ik} variables.

$$(4) \quad f(Y_i)^1 = b_o + \sum_{j=1}^{7} b_j \, f(W_{ij})^1 + e_i$$

The Stepwise Characteristics of the Regression Analysis

With respect to the computational tasks for all of the regression analyses to be performed in this study, the stepwise regression analysis method given by the Biomedical Computer Program[2] will be adapted. This method is based upon the least-squares principle, a sequence of multiple linear regression equations is computed in a stepwise manner. At each

---

[1] $f(Y_i)$ and $f(W_{ij})$ are respectively the transformations of the total cost and the component weights such as for example, the square roots, the reciprocals, the $\log_e$ values, the $\log_{10}$ values etc.

[2] W. J. Dixon. Biomedical Computer Programs (Berkeley, California. University of California Press, 1968), p. 233.

step one variable is added to the regression equation. The variable added is the one which makes the greatest reduction in the error sum of squares.

The following symbols which are related to regression analysis will be used throughout this study.

Y = independent variables.

$\hat{Y}$ = predicted value of the independent variable.

$\bar{Y}$ = sample mean of the independent variable.

σ_Y = sample standard deviation of the independent variable.

v = coefficient of variation of the independent variable, is equal to $\frac{\sigma_Y}{\bar{Y}}$.

R = multiple correlation coefficient. ($R = r$ if there is only one independent variable; r is the correlation coefficient).

s = standard error of estimate.

b = regression coefficient.

σ_b = standard error of the regression coefficient.

n = number of observations.

p = number of independent variables.

e = residual (the error term).

Nature and Limitation of the Available Data With Respect to the Statistical Analyses to be Performed

Cost Data

The data in the cost file are in reality the estimated cost of the contractors, but in the context of the statistical analyses to be performed in this study, they are not estimated data; they are the given observations. These data will not be referred to as estimated data for the statistical analyses so as to avoid confusion with the estimated entities of the regression equations.

Weight Data

The component weights of the ships are given in the weight file in tons. As mentioned earlier, the ship classes found under each type of ship in the weight file are not exactly the same as those found in the cost file. (Recall that each class represents a prototype of ship. One type of ship may include one or more different classes. There are 59 different classes in the cost file and 143 classes in the weight file). This means that the true weight for the ship's components are not available for all the ships in the cost file. Weight averages were computed for the component weights for all the classes under a given ship type in the weight file and these computed average component weights were assigned to every ship under the same type in the cost file. Use of approximate weights and not the true weight will affect the accuracy of the experiments.

There are some ship types in the cost file for which weight information is unavailable. Therefore, in the cost file, the number of records must be reduced from 989 records to 928 records and the number of types of ships must be reduced from 29 to 26. Detailed descriptions of the 26 different types of ships which are to be used for statistical analyses are shown in Table 1.

The following abbreviations will be used for the ship component weights:

HW - Hull structure weight

PW - Propulsion weight

EW - Electric plant weight

CW - Communication and control equipment weight

AUW - Auxiliary system weight

OFW - Outfit and furnish weight

ARW - Armament weight

Price Deflator

The data in the cost file have been collected from the years 1954 to 1966. The statistical analysis aims at serving the engineering approach where only cost as a function of weight is relevant. The change in price which is a function of time is irrelevant and should be eliminated to avoid introduction of additional error variance.

Therefore, all cost data in the cost file were adjusted with the price deflator for producer's durable equipment obtained from the Economic Report of the President, 1969.

TABLE 1

DESCRIPTION OF SHIP TYPES

Type	Description	Number of Classes Under This Type	Number of Ships Under This Type
AE	Ammunition ship	3	33
AFS	Combat store ship	1	24
AGOR	Research ship	3	62
AGS	Surveying ship	2	13
AKA	Cargo ship	1	18
AOE	Fast Combat Support ship	1	4
AOR	Replenishment oiler	1	8
AS	Submarine tender	3	19
CVA	Attack aircraft carrier	3	8
DD	Destroyer	2	19
DDG	Guided missile destroyer	1	109
DE	Escort ship	5	126
DEG	Guided missile escort ship	1	27
DLG	Guided missile frigate	3	74
DLGN	Guided missile frigate	1	3
LPD	Amphibious transport dock	2	29
LPH	Amphibious assault ship	1	10
LSD	Dock landing ship	2	36
LST	Tank landing ship	2	26
MSC	Mine sweeper	3	106
MSI	Mine Sweeper	2	22

TABLE 1 -- Continued

Type	Description	Number of Classes Under This Type	Number of Ships Under This Type
PGM	Mine sweeper	1	22
SS	Submarine	2	8
SSBN	Ballistic missile submarine	4	31
SSN	Nuclear submarine	5	63
YP	Patrol craft	1	28

Data Processing

Because of the large volume of data involved, all operations on the data and computations must be performed by the computer, except for some simple calculations of ratios which were done by slide rule.

The data processing phase constitutes a large amount of effort in this study. The cost file was not arranged in formats which could be used readily as data input for performing the needed computations with the computer, and the file had to be edited for errors and missing data.[1]

Numerous computer programs have to be written to perform the task of analyzing the contents, detecting errors and correcting errors in the cost file. Another important task is to write the computer programs to create numerous input files which contain data from both the cost file and weight file and put in format which allows computer execution of the numerous varieties of regression analyses and other computations.

---

[1]The author wishes to express his appreciation to Mr. Jason Benderly for his joint effort with the editing task.

CHAPTER II

MULTIPLE REGRESSION ANALYSIS

The purpose of this chapter is to answer the basic question of whether it is possible to derive multiple regression equations to predict the total cost of a ship if the component weights are known.

Using the Total 928 Ships

A stepwise regression analysis was performed using the total cost of the ship as the dependent variable and the seven component weights (HW, PW, CW, AUW, EW, OFW, ARW) as independent variables. The entire 928 ships were used. The results are summarized in Table 2.

TABLE 2

RESULTS OF STEPWISE LINEAR REGRESSION ANALYSIS FOR THE TOTAL 928 SHIPS. TOTAL COST AS DEPENDENT VARIABLE, COMPONENT WEIGHTS[1] AS INDEPENDENT VARIABLES. n = 928

Independent variable added	R^2	ΔR^2	$s/\bar{Y}$
PW	.2929	.2929	1.52
HW	.3111	.0182	1.50
AUW	.3165	.0054	1.49
OFW	.3190	.0025	1.49
EW	.3214	.0025	1.485
CW	.3274	.0059	1.48
ARW	.3274	.0000	1.48

$\bar{Y}$ = 176,170 σ_Y = 317,977 v = 1.8

[1] "Component weights" should always mean the seven component weights HW, PW, EW, CW, AUW, OFW, ARW unless otherwise specified.

The following discussion is intended to familiarize the readers with the interpretation of the results of the stepwise regression analyses presented. As mentioned earlier, the stepwise regression analysis performs the computation by steps, adding independent variables to the multiple regression equation one at a time. The variable which contributes most to the reduction of the error sum of squares (or equivalently contributes most to increase the value of R^2) is the one to be included ahead of the other remainders. For example, in reading Table 2, PW is the one which has the greatest contribution, HW comes next, AUW comes third and so on. In some cases, as should be seen in the later experiments, when an independent variable is added, the increase in R^2 is so small as to be negligible, the computation will stop and ignore all remaining independent variables because it is useless to make any further computations. In reading the tables, the following notations should be followed:

$\bar{Y}$ = Sample mean of the dependent variable.

σ_Y = Sample standard deviation of the dependent variable.

v = Coefficient of variation.

$$v = \frac{\sigma_Y}{\bar{Y}}$$

s = Standard error of estimate.

$$s = \left[\frac{1}{n - p - 1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \right]^{1/2}$$

ΔR^2 = Increase in R^2.

One important caution in reading the results in the table is that when an independent variable added shows little contribution to increasing the value of R^2, one should not immediately assume that this independent variable is unrelated to the dependent variable. The fact is that sometimes the independent variables are highly correlated among themselves, they share the same characteristics that contribute to the value of R^2, so that when the contribution to R^2 is assigned to the first independent variable to enter the equation, it will not be repeatedly assigned to the independent variables that come later. To prove and demonstrate this point, seven separate two-variable regression analyses were performed using the total 928 records, with total cost of the ship as dependent variable and with one and only one of the component weights as independent variable. The results of the seven separate two-variable regression analyses were tabulated in Table 3.

TABLE 3

RESULTS OF TWO-VARIABLE REGRESSION ANALYSIS FOR THE TOTAL 928 SHIPS. TOTAL COST AS DEPENDENT VARIABLE, COMPONENT WEIGHT AS INDEPENDENT VARIABLE. n = 928

Independent variable	R^2	ΔR^2	$s/\bar{Y}$
HW	.2530	.2530	1.56
PW	.2929	.2929	1.51
EW	.2878	.2878	1.53
CW	.1430	.1430	1.67
AUW	.2430	.2430	1.57

TABLE 3 -- Continued

Independent variable	R^2	ΔR^2	$s/\bar{Y}$
OFW	.1432	.1432	1.67
ARW	.1000	.1000	1.70

$\bar{Y} = 176,170$ $\sigma_Y = 317,977$ $v = 1.8$

The values of ΔR^2 for the weight components in Table 3 are much higher than those in Table 2 except for PW which was the first independent variable to enter in Table 2. Looking at the results of Table 2 and Table 3, it is evident that the values of R^2 and $s/\bar{Y}$ in these experiments are not satisfactory. One visible reason is the value of the standard deviation of the dependent variable σ_Y, which has a high value and consequently caused the value of the coefficient of variation v to be high, this is an indication that the sample is highly dispersed relative to the mean.

At this point, it seems necessary to clarify what is to be considered satisfactory or not satisfactory for the value of R^2 and $s/\bar{Y}$. The judgement to be made here must necessarily be subjective. From the characteristics of the numerous experiments which one will observe later, it appears that an R^2 of less than .70 and/or a $s/\bar{Y}$ of greater than .30 are usually associated with a regression analysis where the sample stratification and/or the regression form used were poor; poor in the sense that much better selections of one or both of these two factors can lead to much improvement in the value of R^2 and $s/\bar{Y}$. For example, one can observe in Table 4 that for the small hull group, the value of R^2 is .1183

and the value of $s/\bar{Y}$ is 2.73 these results are evidently poor. By further stratification of this small hull group, it is possible to improve the value of R^2 to .8135 and $s/\bar{Y}$ to .24 as obtained in the Small Hull Subgroup-1 shown in Table 7. Therefore, from the viewpoint of statistical significance, an R^2 of less than .70 and/or an $s/\bar{Y}$ of greater than .30 will be arbitrarily considered as unsatisfactory. But when it comes to the practical purpose of predicting the total cost of the ship (which is the dependent variable of the regression analysis) an $s/\bar{Y}$ value of greater than .05 can still be judged as unsatisfactory if one considers that in most of the ships under study, the total cost of the ship is often much over ten million dollars. In this case, the determining factor which is important is the purpose for which the predicted value is to be used by the decision maker.

It should be noted that the variable $s/\bar{Y}$ is favored to be used over the value of s in this study because it enables comparisons between different experiments where samples are made up of ship groups with different magnitudes of the dependent variable. The use of $s/\bar{Y}$ provides a common unit for comparing results of the different regression models.

Stratification by Hull Weight

Hoping that the regression analyses results would improve if the samples used were more homogeneous, the total 928 ships were stratified into 3 groups with respect to magnitude of hull weight.

(1) Small Hull Group - Ships with hull weights of less than 2000 tons. There are 12 types[1] with 605 ships in total.

---

[1] The types are: AGOR, AGS, DD, DDG, DE, DEG, MSC, MSI, PGM, SS, SSN, YP.[2]

[2] Detailed descriptions of ship types are given in Table 1.

(2) Medium Hull Group - Ships with hull weights between 2000 and 7000 tons, there are 11 types[1] with 303 ships in total.

(3) Large Hull Group - Ships with hull weights above 8000 tons. There are 3 types[2] with 20 ships in total.

Three stepwise regression analyses were performed for the above three groups using the total cost of the ship as dependent variable and the seven component weights as the independent variables. The results are shown in Tables 4, 5, and 6.

The results show significant improvement for the Medium Hull group and the Large Hull group. The value of R^2's had increased and the value of $\frac{s}{\bar{Y}}$ had decreased substantially from those in Table 2. It should be observed that these two groups had a relatively small value for the coefficient of variation v . However, there was not improvement for the Small Hull group, it can be seen that this group has a high coefficient of variation, v = 1.8 .

---

[1]The types are: AE, AFS, AKA, AS, DLG, DLGN, DPD, LPH, LSD, LST, SSBN.[3]

[2]The types are: AOE, AOR, CVA.[3]

[3]Detailed descriptions of ship types are given in Table 1.

TABLE 4

RESULTS OF STEPWISE LINEAR REGRESSION ANALYSIS FOR THE SMALL HULL GROUP WHICH CONSISTS OF SHIPS WITH HULL WEIGHT LESS THAN 2000 TONS. TOTAL COST AS DEPENDENT VARIABLE, COMPONENT WEIGHTS AS INDEPENDENT VARIABLES.

$n = 605$

Independent Variable added	R^2	ΔR^2	$s/\bar{Y}$
HW	.0956	.0956	2.76
OFW	.1127	.0171	2.73
EW	.1180	.0054	2.72
PW	.1183	.0002	2.73

$\bar{Y} = 115{,}557$

$\sigma_Y = 334{,}630$

$v = 2.9$

TABLE 5

RESULTS OF STEPWISE LINEAR REGRESSION ANALYSIS FOR THE MEDIUM HULL GROUP WHICH CONSISTS OF SHIPS WITH HULL WEIGHT FROM 2000 to 7000 TONS. TOTAL COST AS DEPENDENT VARIABLE, COMPONENT WEIGHTS AS INDEPENDENT VARIABLES.

$n = 303$

Independent variable added	R^2	ΔR^2	$s/\bar{Y}$
EW	.2308	.2308	.330
OFW	.3424	.1116	.306
ARW	.3725	.0302	.299
HW	.3941	.0215	.294
AUW	.4274	.0330	.285
CW	.4290	.0016	.286
PW	.4327	.0057	.285

$\bar{Y} = 250{,}081$

$\sigma_Y = 93{,}704$

$v = .376$

TABLE 6

RESULTS OF STEPWISE LINEAR REGRESSION ANALYSIS FOR THE LARGE HULL GROUP WHICH CONSISTS OF SHIPS WITH HULL WEIGHT GREATER THAN 7000 TONS. TOTAL COST AS DEPENDENT VARIABLE, COMPONENT WEIGHT AS INDEPENDENT VARIABLES.

$n = 20$

Independent variable added	R^2	ΔR^2	$s/\bar{Y}$
AUW	.9386	.9386	.201
CW	.9388	.0002	.205

$\bar{Y} = 889{,}961$

$\sigma_Y = 702{,}641$

$v = .795$

The results show significant improvement for the Medium Hull group and the Large Hull group. The value of R^2's had increased and the value of $\frac{s}{\bar{Y}}$ had decreased substantially from those in Table 2. It should be observed that these two groups had a relatively small value for the coefficient of variation v . However, there was no improvement for the Small Hull group, it can be seen that this group has a high coefficient of variation, $v = 1.8$.

It seems reasonable to make the following postulates at this point:

Postulate I — A good result[1] from regression analysis can not be expected if the coefficient of variation of the dependent variable v is high.

Postulate II — That the result of the regression analysis can be improved if the sample can be stratified into subsamples which have smaller value of v .

<u>Further Stratification of the Hull Weight</u>

To test Postulates I and II just mentioned in the previous section, the small hull group whose results were reflected in Table 4 to be poor was further stratified into three subgroups:

(1) Small Hull Subgroup-1 - Ships with hull weights less than 500 tons. There are 4 types[2] with a total of 128 ships.

---

[1]A good result in regression analysis means here a low value of $\frac{s}{\bar{Y}}$ and a high value of R^2.

[2]The types are: MSC, MSI, PGM, YP.[3]

[3]Detailed description of ship types are given in Table 1.

(2) Small Hull Subgroup-2 - Ships with hull weight between 500 to 1000 tons. There are 5 types[1] with a total of 228 ships.

(3) Small Hull Subgroup-3 - Ships with hull weights between 1000 to 2000 tons. There are 3 types[2] with a total of 199 ships.

Stepwise regression analyses were performed for these three groups with total cost as dependent variable and the component weights as independent variables. The results are shown in Tables 7, 8 and 9.

The result shows significant improvement of the Subgroup-1 and the Subgroup-2, but no improvement for the Subgroup-3. Looking at the value of v, it is .55 for the Subgroup-1, .545 for the Subgroup-2, and 2.38 for the Subgroup-3.

This reinforces Postulates I and II which stated that it is impossible to obtain good regression results if v is large and that results can be improved if the value of v can be reduced by proper stratification. However, a new postulate can be made here, that is:

Postulate III - Stratification into smaller sub-groups does not always result in reduction in the value of v.

---

[1]The types are: AGOR, AGS, DD, DE, SS.[3]

[2]The types are: DDG, DEG, SSN.[3]

[3]Detailed descriptions of ship types are given in Table 1.

TABLE 7

RESULTS OF STEPWISE LINEAR REGRESSION ANALYSIS FOR THE SMALL HULL SUBGROUP-1 WHICH CONSISTS OF SHIPS WITH HULL WEIGHT LESS THAN 500 TONS. TOTAL COST AS DEPENDENT VARIABLES, COMPONENT WEIGHTS AS INDEPENDENT VARIABLES.

$n = 178$

Independent variable added	R^2	ΔR^2	$s/\bar{Y}$
AUW	.8092	.8092	.24
EW	.8135	.0043	.24

$\bar{Y} = 12{,}523$

$\sigma_Y = 6{,}932$

$v = .55$

TABLE 8

RESULTS OF STEPWISE LINEAR REGRESSION ANALYSIS FOR THE SMALL HULL SUBGROUP-2 WHICH CONSISTS OF SHIPS WITH HULL WEIGHT BETWEEN 500 to 1000 TONS. TOTAL COST AS DEPENDENT VARIABLE, COMPONENT WEIGHTS AS INDEPENDENT VARIABLES.

$n = 228$

Independent variable added	R^2	ΔR^2	$s/\bar{Y}$
PW	.8397	.8397	.220
HW	.8838	.0441	.187
AUW	.9054	.0216	.169
EW	.9060	.0006	.169

$\bar{Y} = 91,375$

$\sigma_Y = 49,952$

$v = .545$

TABLE 9

RESULTS OF STEPWISE LINEAR REGRESSION ANALYSIS FOR THE SMALL HULL SUBGROUP-3 WHICH CONSISTS OF SHIPS WITH HULL WEIGHT BETWEEN 1000 to 2000 TONS. TOTAL COST AS DEPENDENT VARIABLE, COMPONENT WEIGHTS AS INDEPENDENT VARIABLES.

$n = 199$

Independent variable added	R^2	ΔR^2	$s/\bar{Y}$
AUW	.0423	.0423	2.33
HW	.0428	.0005	2.34

$\bar{Y} = 235,432$

$\sigma_Y = 560,374$

$v = 2.38$

Stratification by Propulsion Weight

At this point, it seems correct that proper stratification results in multiple regression equations which have a higher value of R^2 and a lower value of $s/\bar{Y}$. But so far all previous attempts show that the value of $s/\bar{Y}$ has not been less than 20%. This is not satisfactory for practical prediction purposes. The question arises as to whether there are some better approaches to stratify the samples other than by hull weight. In this section, the results of stratification by propulsion weight will be discussed. The entire 928 ships were divided on the basis of propulsion weight into:

(1) Small Propulsion - Ships with propulsion weight less than 500 tons. There are 10 types[1] which make a total of 440 ships.

(2) Medium Propulsion - Ships with propulsion weight between 500 and 1000 tons. There are 12 types[2] with 465 ships in total.

(3) Large Propulsion - Ships with propulsion weight greater than 1000 tons. There are 4 types[3] with 23 ships in total.

Stepwise regression analysis was performed as before and the results are shown in Table 10, 11 and 12. The results show no significant improvement over the stratification by hull weight, it exhibits the same characteristics that a low value of v leads to better results and a high value of v can lead to very poor results.

---

[1]The types are: AGOR, AGS, DE, DEG, LST, MSC, MSI, PGM, SS, YP.[4]

[2]The types are: AE, AFS, AKA, AS, DD, DDG, DCG, LPD, LPH, LSD, SSBN, SSN.[4]

[3]The types are: AOE, AOR, CVA, DLGN.[4]

[4]Detailed descriptions of ship types are given in Table 1.

TABLE 10

RESULTS OF STEPWISE LINEAR REGRESSION ANALYSIS FOR THE SMALL PROPULSION GROUP WHICH CONSIST OF SHIPS WITH PROPULSION WEIGHT LESS THAN 500 TONS. TOTAL COST AS DEPENDENT VARIABLE, COMPONENT WEIGHTS AS INDEPENDENT VARIABLES.

$n = 440$

Independent variable added	R^2	ΔR^2	$s/\bar{Y}$
PW	.8846	.8846	.288
HW	.8952	.0107	.273
EW	.8991	.0039	.269
OFW	.9078	.0087	.258
ARW	.9136	.0057	.249
CW	.9179	.0043	.243
AUW	.9187	.0008	.242

$\bar{Y} = 57,883$

$\sigma_Y = 48,600$

$v = .84$

TABLE 11

RESULTS OF STEPWISE LINEAR REGRESSION ANALYSIS FOR THE MEDIUM PROPULSION GROUP WHICH CONSISTS OF SHIPS WITH PROPULSION WEIGHT BETWEEN 500 to 1000 TONS. TOTAL COST AS DEPENDENT VARIABLE, COMPONENT WEIGHTS AS INDEPENDENT VARIABLES.

$n = 465$

Independent variable added	R^2	ΔR^2	$s/\bar{Y}$
EW	.0033	.0033	1.45
OFW	.0131	.0099	1.44
CW	.0295	.0163	1.43
AUW	.0344	.0049	1.43
PW	.0348	.0004	1.43
HW	.0353	.0005	1.44
ARW	.0359	.0006	1.44

$\bar{Y} = 254{,}717$

$\sigma_Y = 370{,}110$

$v = 1.46$

TABLE 12

RESULTS OF STEPWISE LINEAR REGRESSION ANALYSIS FOR THE LARGE PROPULSION GROUP WHICH CONSISTS OF SHIPS WITH PROPULSION WEIGHT GREATER THAN 1000 TONS. TOTAL COST AS DEPENDENT VARIABLE, COMPONENT WEIGHTS AS INDEPENDENT VARIABLES.

$n = 23$

Independent variable added	R^2	ΔR^2	$s/\bar{Y}$
ARW	.9175	.9175	.228
OFW	.9273	.0097	.220
PW	.9382	.0110	.206

$\bar{Y} = 851,066$

$\sigma_Y = 661,724$

$v = .78$

Stratification by Usage

Not satisfied with the hull weight and propulsion weight stratification, three groups of ships having the same usage were selected for analyses.

(1) Transport Ship group. Includes type AE, AFS, AS, AKA. The total number of ships is 94.

(2) Combat Ship group. Includes type DD, DDG, DE, DEG, DLG, DLGN. The total number of ships is 358.

(3) Amphibious Ship group. Includes types LPD, LPH, LSD, LST. The total number of ships is 101.

Three stepwise regression analyses were performed for these groups with total cost as dependent variable and the component weights as independent variables. The results are shown in Tables 13, 14 and 15.

The result shows convincing improvement over the previous two stratifications in the following two aspects. First, it seems more unlikely in this stratification scheme to come up with an extremely large value of v thus very poor results and second, the value of $\frac{s}{\bar{Y}}$ has been reduced and at some instance to as low as 13.3%, which the previous two stratifications were not able to achieve. However, a $\frac{s}{\bar{Y}}$ value of 13.3% is still not satisfactory for practical prediction.

Analyses of Ratios and Coefficients

In order to facilitate overall analysis of the various experiments just performed, two tables were prepared. Table 16 tabulates the values of v , $s/\bar{Y}$ and R^2 for the 13 different groups. Table 17 tabulates the correlation coefficient, r_{xy} , between the total cost (dependent variable) and the component weights (independent variables) for the 13 different groups.

TABLE 13

RESULTS OF STEPWISE LINEAR REGRESSION ANALYSIS FOR THE TRANSPORT SHIP GROUP. TOTAL COST AS DEPENDENT VARIABLE, COMPONENT WEIGHTS AS INDEPENDENT VARIABLES.

$n = 94$

Independent variable added	R^2	ΔR^2	$s/\bar{Y}$
OFW	.7156	.7156	.174
ARW	.7162	.0007	.175

$\bar{Y} = 221{,}420$

$\sigma_Y = 71{,}662$

$v = .323$

TABLE 14

RESULTS OF STEPWISE LINEAR REGRESSION ANALYSIS FOR THE <u>COMBAT SHIP GROUP</u>. TOTAL COST AS DEPENDENT VARIABLE, COMPONENT WEIGHTS AS INDEPENDENT VARIABLES.

$n = 358$

Independent variable added	R^2	ΔR^2	$s/\bar{Y}$
EW	.8193	.8193	.180
PW	.8790	.0596	.146
CW	.8986	.0196	.135
OFW	.8997	.0011	.134
ARW	.9022	.0025	.133

$\bar{Y} = 161{,}531$

$\sigma_Y = 68{,}085$

$v = .425$

TABLE 15

RESULTS OF STEPWISE LINEAR REGRESSION ANALYSIS FOR THE AMPHIBIOUS SHIP GROUP. TOTAL COST AS DEPENDENT VARIABLE, COMPONENT WEIGHTS AS INDEPENDENT VARIABLES.

n = 101

Independent variable added	R^2	ΔR^2	$s/\bar{Y}$
CW	.7899	.7899	.160
HW	.8218	.0319	.148
ARW	.8221	.0003	.148

$\bar{Y}$ = 226,419

σ_Y = 78,586

v = .35

TABLE 16

SUMMARY TABLE FOR THE VALUES OF v, $s/\bar{Y}$ AND R^2 OF THE STEPWISE LINEAR REGRESSION ANALYSES FOR THE DIFFERENT STRATIFICATIONS.

Stratification	v	$s/\bar{Y}$	R^2
Total 928 ships	1.800	1.480	.3274
Small Hull group	2.900	2.730	.1183
Medium Hull group	0.376	0.285	.4327
Large Hull group	0.795	0.205	.9388
Small Hull Subgroup-1	0.550	0.240	.8135
Small Hull Subgroup-2	0.545	0.169	.9060
Small Hull Subgroup-3	2.380	2.340	.0428
Small Propulsion group	0.840	0.242	.9187
Medium Propulsion group	1.460	1.440	.0359
Large Propulsion group	0.780	0.206	.9382
Transport Ship group	0.323	0.175	.7162
Combat Ship group	0.425	0.133	.9022
Amphibious Ship group	0.350	0.148	.8221

TABLE 17

CORRELATION COEFFICIENTS r_{xy} BETWEEN THE DEPENDENT VARIABLE AND THE INDEPENDENT VARIABLES FOR THE DIFFERENT STRATIFICATIONS.

Stratifications	HW	PW	EW	CW	AUW	OFW	ARW
Total 928 ships	.50	.54	.54	.38	.49	.38	.32
Small Hull group	.31	.30	.30	.19	.26	.15	.17
Medium Hull group	.03	.41	.48	.34	-.11	.02	.19
Large Hull group	.97	.97	.96	.97	.97	.97	.96
Small Hull Subgroup-1	.90	.31	.88	.90	.90	.89	.12
Small Hull Subgroup-2	.81	.92	.82	.88	.59	.70	.91
Small Hull Subgroup-3	.20	.11	.20	-.21	.21	-.21	-.14
Small Propulsion group	.83	.94	.89	.63	.77	.82	.88
Medium Propulsion group	.01	.02	.06	-.04	-.02	-.02	-.01
Large Propulsion group	.93	.95	.96	.72	.94	.92	.96
Transport Ship group	.67	-.80	.76	-.15	.12	.85	-.13
Combat Ship group	.80	.91	.91	.86	.79	.83	.82
Amphibious Ship group	.85	.58	.81	.89	.85	.85	.61

Note:

The entries in the table are value of r_{xy}'s between the total cost of the ship and the respective component weights which are shown at the top of the column.

Observation of Table 16 confirms again the earlier postulate that it is impossible to obtain good regression results if the value of the coefficient of variation, v , is high. In every case where the value v is over 1.00 the result is extremely poor. The value of $\frac{s}{\bar{Y}}$ tends to increase or decrease with v. However, the value of R^2 is also a determining factor. A relatively low $\frac{s}{\bar{y}}$ can be obtained for a relatively high v if R^2 is high, this is shown in the Large Hull group. Conversely, a relatively high $\frac{s}{\bar{Y}}$ can occur for a relatively low v if R^2 is low. This is shown in the Medium Hull group. All these point to the fact that $\frac{s}{\bar{Y}}$ is a function of both v and R^2 . It increases with v and decreases with R^2 .

Observing that both Tables 16 and 17 indicate the following relationships:

(1) In a particular group, if all the r_{xy}'s are high, R^2 is also high.

(2) In a particular group, if all the r_{xy}'s are low, R^2 is also low.

(3) In a particular group, if some of the r_{xy}'s are high and some are low, it is possible that R^2 can still be high. This is shown in the Transport Ship group.

Observation of Table 17 indicates the following relationships:

(1) There is not any component weight that shows consistently higher values of correlation with the dependent variable over the other component weights throughout all the different experiments.

(2) The component weights CW, AUW, OFW, ARW have more occurrence of negative r_{xy} than HW, PW, and EW.

(3) In a particular group, a low value of r_{xy} for all the independent variables does not always mean that there is a lack of true correlation between the dependent and independent variables. With the present ship cost data, it would seem more likely that it is due to poor stratification of samples. This is shown in the Small Hull Group where r_{xy}'s are small, but after further stratification, the Small Hull Subgroup-1 and Small Hull Subgroup-2 produced high values of r_{xy}'s.

Correlation Coefficients Between the Independent Variables

The correlation matrix for the independent variable was computed for each experiment performed. It was found that in many cases the correlation coefficients between the independent variables, r_{xx}'s , are high, i.e., almost as large as the correlation coefficients between the dependent and independent variables, r_{xy}'s . In several of the regression analyses just performed, it was found that if the independent variable which accounted for the highest contribution to R^2 was eliminated from the computation, some other independent variable would come up in its place, and as a result, the value of R^2 did not change significantly. This high value of r_{xx}'s signifies that the problem of multicollinearity does exist.

Zero Regression Intercept

In this section, multiple regression analysis which calls for a zero regression intercept will be used. Under this method, b_o , the constant term of the regression equation will be zero. The covariances, standard deviations and correlations are computed about the origin rather than about the mean.

The method was performed on the total 928 ships. The following is a comparison of the results:

	Best[1] R^2 obtained	Best[2] s obtained	Best $s/\bar{Y}$
zero intercept not called for	.3274	261,779	1.480
zero intercept called for	.4840	262,019	1.485

It can be observed that although the value of R^2 was improved, there was no improvement on the value of s , which was not surprising.

[1]Best R^2 means the largest value of R^2 obtained from the stepwise regression analysis.

[2]Best s means the smallest value of standard error of estimate obtained from the stepwise regression analysis.

The same experiment was performed on the Amphibious Ship group. The Amphibious Ship group shows consistently better value of R^2 and s among the many groups chosen for regression analysis in this chapter. It is therefore favored for testing new methods to find out if the new method can make further improvement for groups which have comparatively good results already. The experiment shows improvement in R^2 and slight improvement in s as follows:

	Best R^2 obtained	Best s obtained	Best $s/\bar{Y}$
zero intercept not called for	.7163	38,802	.175
zero intercept called for	.9811	33,655	.152

Above two experiments seem to indicate that better regression equations may be obtained by calling for zero regression intercept. However, the improvements may not be significant because there may be no improvement or only slight improvement in the value of the standard error of estimate.

SUMMARY

The following is a summary of the principal findings of the regression analyses thus far described.

(1) The parameters v R^2, $s/\bar{Y}$ are important in examining the results of the regression analyses. The value of $s/\bar{Y}$ is a function of v and R^2, it increases with v and decreases with R^2.

(2) If the value of all the correlation coefficients between the dependent and independent variables r_{xy}'s are high, R^2 is likely to be high. If all r_{xy}'s are low, R^2 is likely to be low. If some r_{xy}'s are high and some are low, R^2 could possibly still be high.

(3) If all the values of r_{xy}'s are found to be low, most likely it is due to poor stratification of samples rather than due to lack of relationships between the dependent and independent variables.

(4) To obtain a good regression result, two basic requirements must be met.

(a) The value of v must be sufficiently low (not higher than 0.8).

(b) The value of the r_{xy} for some of the independent variables, but not necessarily all the independent variables, must be sufficiently high (higher than .75).

These requirements can be fairly satisfied by proper stratification of samples.

(5) In selecting the samples, it seems grouping by usage is more promising than grouping by hull weight or grouping by propulsion weight.

(6) The results of regression analysis may be slightly improved by using a zero regression intercept.

(7) Of all the various regression analyses performed in this chapter, the lowest value of $s/\bar{Y}$ is 13.3% which is still not satisfactory for practical purposes because the total cost of a ship is usually in the magnitude of many million dollars such that the tolerance for error is small. It is suggested that better results could be achieved if the actual weight instead of the approximate weight of the ship components were used, but how significant will be the improvement is still unknown.

CHAPTER III

THE ADDING-UP PROCESS

Adding-up Process for the Total 928 Ships

Using the total 928 ships, seven sets of two-variable regression analyses were performed with component weight as the independent variable and the component cost as the dependent variable, i.e., hull cost versus hull weight, propulsion cost versus propulsion weight, etc. The regression equation takes the form

$$C_{ij} = b_{oj} + b_j W_{ij} + e_i \,, \qquad \begin{matrix} i = 1 \text{ to } 928 \\ j = 1 \text{ to } 7 \end{matrix} \tag{4-1}$$

where C_{ij} is the component cost

W_{ij} is the component weight

e_i is the residual .

After the parameters b_{oj}'s and b_j's have been determined, the estimated total cost of the ship was computed by the equation

$$\hat{Y}_i = \sum_{j=1}^{7}(\hat{b}_{oj} + \hat{b}_j W_{ij}) + CSC_i + DEC_i + Profit_i \qquad i = 1 \text{ to } 928 \tag{4-2}$$

where CSC_i is the construction cost

DEC_i is the design and engineer cost.

The standard error of estimate was computed with the equation

$$s = \left[\frac{1}{928-10-1} \sum_{i=1}^{928} (Y_i - \hat{Y}_i) \right]^{\frac{1}{2}} \tag{4-3}$$

The value of s computed here shows significant improvement over the one obtained by the multiple regression analysis.

	s	$\frac{s}{\bar{Y}}$
Adding-up process	50,350	0.288
Multiple regression analysis	261,779	1.480

II. <u>Adding-up Process for the Amphibious Ship Group</u>

Due to the remarkable improvement of the above experiment, the question of interest becomes whether the adding-up process will also show improvement for ship groups whose multiple regression results are already comparatively good such as the Amphibious Ship group.

The same procedures as in equations (4-1) to (4-3) were applied to the Amphibious Ship group (except in this case $i = 1$ to 101) . The results are compared with those obtained from the multiple regression analysis as follows

	s	$\frac{s}{\bar{Y}}$
Adding-up process	42,000	.186
Multiple regression analysis	33,654	.152

Contradictory to the results obtained with the 528 ships, for the Amphibious Ship group, the adding up process shows poorer results than the multiple regression analysis.

<u>Adding-up Process for the Transport Ship Group</u>

To pursue the investigation further, the same experiment was performed on the Transport Ship group, which also has comparatively good results from the multiple regression analysis. The experiment

here again shows that the adding-up process yields poorer results than the multiple regression analysis as shown below.

	s	$\frac{s}{\bar{Y}}$
Adding-up process	54,892	.248
Multiple regression analysis	38,802	.175

SUMMARY

The standard error of estimate for estimating the total cost of the ship was compared between the adding-up process and the multiple regression analysis. It was found that the adding-up process yields better results for the 928 ships but poorer results for the Amphibious and Transport Ship groups.

It should be noted that the 928 ships can be considered as a poorly selected sample due to its high value of coefficient of variation v, its multiple regression analysis results are poorer than the other two groups as shown below.

	Coefficient of variation v	Multiple regression R	$\frac{s}{\bar{Y}}$
928 ships	1.800	.5721	1.480
Amphibious Ship group	0.323	.8463	0.152
Transport Ship group	0.350	.9067	0.175

For the 928 ships, the values of the R's obtained from the two-variable regression analysis of the adding-up process as represented by equation (4-1) are much higher in value than the value of R obtained from the multiple regression analysis. However, this is not true for the Amphibious Ship group and the Transport Ship group where the multiple regression R is high.

The above facts seem to lead to the conclusion that for poorly selected samples, the adding-up process will yield better results than the multiple regression analysis. Whereas for better selected samples, the multiple regression analysis will yield better results than the adding-up process. Hence multiple regression analysis should be preferred.

CHAPTER IV

NON-LINEAR REGRESSION ANALYSIS

All the previous regression analyses performed assumed linear relationships. The experiments which are to be performed here aim at testing whether there are some non-linear relationships existing between the dependent and independent variables.

Transforming the Independent Variables

Using the total 928 ships, stepwise regression analyses were performed with total cost of the ship as the dependent variable and the transformed values of the component weights W_{ij} which we shall call $f(W_{ij})$ as the independent variables. The different transformed values to be tested are W_{ij}^2, $\sqrt{W_{ij}}$, $\frac{1}{W_{ij}}$, $\log_{10} W_{ij}$, W_{ij}^3, $\log_e W_{ij}$.

The regression equation takes the form

$$Y_i = b_o + \sum_{j=1}^{7} b_j \; f(W_{ij}) + e_i, \; i = 1 \text{ to } 928.$$

The results of the regression analyses are shown in Table 18. In this table, each column represents the results obtained from a particular regression analysis with a particular transformation. For example, the column under W_{ij} shows results of regression analysis where the original W_{ij}'s are used as independent variables; the column under W_{ij}^2 shows the results of the regression analysis where the squares of W_{ij}'s are used as independent variables, etc. Examination of Table 18 shows no improvement over the linear model for the different transformations.

TABLE 18

RESULTS OF STEPWISE NON-LINEAR REGRESSION ANALYSIS FOR THE TOTAL 928 SHIPS. TOTAL COST AS DEPENDENT VARIABLE, TRANSFORMED VALUES OF COMPONENT WEIGHTS AS INDEPENDENT VARIABLES

	W_{ij}	W_{ij}^2	$\sqrt{W_{ij}}$	$\log_{10} W_{ij}$	$\frac{1}{W_{ij}}$	W_{ij}^3	$\log_e W_{ij}$
r_{xy} value for HW	.50	.47	.48	.39	-.22	.45	.39
- do - PW	.54	.52	.49	.38	-.25	.47	.37
- do - EW	.54	.50	.48	.38	-.20	.47	.37
- do - CW	.38	.39	.35	.29	-.14	.38	.28
- do - AUW	.49	.46	.46	.36	-.19	.45	.36
- do - OFW	.38	.41	36	.32	-.22	.43	.32
- do - ARW	.32	.27	.35	.34	-.25	.24	.34
Value of R	.57	.54	.55	.46	.32	.49	.45
Value of $\frac{s}{\bar{Y}}$	1.48	1.52	1.52	1.61	1.71	1.57	1.62

$\bar{Y}$ = 176,170

Including More Independent Variables into the Model

Two other forms of non-linear relationships will be tried here:

(1) Using the 928 ships, in addition to the seven component weights, the squares of all the seven component weights are also included as independent variables. The regression equation takes the form

$$Y_i = \sum_{j=1}^{7} b_j W_{ij} + \sum_{j=1}^{7} b_j W_{ij}^2 + e_i \qquad (3\text{-}1)$$

$$i = 1 \text{ to } 928.$$

(2) Using the 928 ships, in addition to the seven component weights, the cross products of the component weights HW, PW, EW, CW were also included as independent variables. The regression equation therefore has the following 18 independent variables.

(a) The seven component weights

(b) HW x PW, PW x EW, EW x CW, HW x CW, HW x EW, PW x CW

(c) HW x PW x EW, HW x PW x CW, PW x EW x CW, HW x PW x CW

(d) HW x PW x CW x EW

The results of the two experiments are shown and compared to the simple linear model below.

	Best R obtained	Best s obtained
Simple linear model	5721	261,779
Using equation (3-1)	6957	262,019
Using the 18 variables in [2]	5763	261,423

The results show no improvements. It was observed that the independent variables which were added to supplement the seven component weights were mostly highly correlated with one or more of the seven component weights. As a consequence, it would not be expected that these variables would produce an improved model.

Transforming the Dependent Variable

Using the total 928 ships, stepwise regression analyses were performed using the seven component weights as independent variables and the transformed values of the total cost Y_i, which we shall call $f(Y_i)$, as the dependent variable. The regression equation takes the form

$$f(Y_i) = b_o + \sum_{j=1}^{7} b_j W_{ij} + e_i \qquad i = 1 \text{ to } 928.$$

The different transformations of Y_i are Y_i^2, $\sqrt{Y_i}$, $\log_e Y_i$.
The results of these various regression analyses are shown in Table 19.

TABLE 19

RESULTS OF STEPWISE NON-LINEAR REGRESSION ANALYSIS FOR THE TOTAL 928 SHIPS. TRANSFORMATION OF TOTAL COST AS DEPENDENT VARIABLE, COMPONENT WEIGHTS AS INDEPENDENT VARIABLES

$n = 928$

	Y_i	Y_i^2	$\sqrt{Y_i}$	$\frac{1}{Y_i}$	$\log_e Y_i$
Best R obtained	.57	.14	.87	.50	.8352
Best s obtained	261,641	*	102	.0001	.7326
$\bar{Y}$	176,170	*	367	.00003	11.45
$\frac{s}{\bar{Y}}$	1.48		.218	33	.0064

*Value extremely large; exceeded size defined in computer program

In Table 19, each column represents the results obtained from a particular regression analysis using a particular transformation of the total cost Y_i as dependent variable.

Table 19 seems to indicate remarkable improvements for the transformation $\sqrt{Y_i}$. There is an increase of the value of R from .57 for the linear model to .87 and a decrease of the value of $\frac{s}{\bar{Y}}$ from 1.48 for the linear model to 0.218. However, to test whether the decrease in the standard error of estimate is mostly due to the change in unit of measurement and whether there is in fact no real improvement, an experiment was performed using the following equations.

$$\widehat{\sqrt{Y_i}} = \hat{b}_o + \sum_{j=1}^{7} \hat{b}_j W_{ij}$$

$\hat{b}_o$ and $\hat{b}_j$ are coefficients obtained from the regression analysis

$$\hat{Y}_i = (\widehat{\sqrt{Y_i}})^2$$

$$e_i = Y_i - \hat{Y}_i \qquad (3\text{-}2)$$

$$s = \left[\frac{1}{928-7-1} \sum_{i=1}^{928} e_i^2 \right]^{\frac{1}{2}} \qquad (3\text{-}3)$$

It was found that the computed s is 268, 113--slightly larger than the value of s obtained by the linear model which is 261,776.

This indicates that results obtained from transformation of variables may be deceptive, it may give an apparent improvement when in fact there is no real improvement.

Computations for the value of s using equations (3.2) and (3-3) were performed also for the regression results where Y_i is transformed to $\log_e Y_i$; in this case , however, $\hat{Y}_i$ is the antilog of the estimated $\log_e Y_i$. The value of s computed here is 1,379,836 which is very much larger than the value of s of the linear model.

Similar experiments as above where Y_i was transformed to $\sqrt{Y_i}$ and $\log_e Y_i$ were performed for the Amphibious Ship group, Combat Ship group and Transport Ship group; all results indicate no improvement in the value of s over the linear model.

<u>Transforming Both the Dependent Variable and the Independent Variables</u>

Using the total 928 ships, stepwise regression analyses were performed with the transformations of the seven component weights as independent variables and the transformation of the total cost as dependent variable. The regression equation becomes

$$f(Y_i) = b_o + \sum_{j=1}^{7} b_j \; f(W_{ij}) + e_i, \qquad i = 1 \text{ to } 928 \;.$$

The different transformations are (1) transforming the variables to their square root, (2) transforming the variables to their $\log_{10}$ value, and (3) transforming the variables to their $\log_e$ value.

The standard error of estimate was also computed for each experiment, using methods similar to the preceding section where the values of $\hat{Y}_i$'s in its non-transformed magnitude were used to compute the value of the residual e's.

The results of these experiments show again improvement in the value of R but no real improvement in the value of s for the different transformations. The value of the R's and the standard error of estimates as described in the preceding paragraph are shown as follows:

	Best R	Best s
Original variables	.5721	261,779
Transforming variables to their square root	.8966	261,767
Transforming variables to their $\log_{10}$ value	.9806	299,032
Transforming variables to their $\log_e$ value	.9806	278,202

Similar experiments as above were performed for the Amphibious Ship group, Combat Ship group and Transport Ship group. The same results in which there is an improvement in the value of R but no real improvement in the value of s are obtained for all these experiments.

Graphical Analysis

To help visualize the data, the Amphibious Ship group was selected for graphical treatment. For each of the seven component weights, a graph was prepared with points plotted between the total cost of the ship and the component weight. Two representative graphs are shown here. Figure 1 is a graph of the total cost versus the hull weight, it represents a well structured relationship between the two variables (the correlation coefficient is .85). Figure 2 is a graph of the total cost versus the propulsion weight, it represents a poor structural relationship between the two variables (the correlation coefficient is .58). For the remaining five component weights whose graphs are not included, those with high correlation coefficients appear similar to Figure 1 and those with low correlation coefficients appear similar to Figure 2.

The appearance of points concentrated in four straight columns in the graph is very informative. It reveals the effect of assigning the same value for each of the component weights to all ships of the same type. The Amphibious Ship group includes four types of ships which consequently produce the four columns in the graph. Because of this existing configuration, no matter where a regression curve is drawn through these points, a substantial number of points deviating from the curve is unavoidable. This could be one good explanation for the high value of the standard errors of estimate (which is a

function of the deviation of the points from the regression curve). This also opens the possibility of making a conjecture that the true inherent relationship between the two variables could possibly be described by a non-linear curve, such as a double-log curve. But the existence of these "point columns" in the present data may render little difference in the computed value of standard error of estimate, whether one fits a linear curve or a non-linear curve through the points as had been encountered in some of the previous experiments. This conjecture can also be applied to the zero regression intercept experiments performed in chapter two.

FIGURE 1

TOTAL COST VERSUS HULL WEIGHT FOR THE AMPHIBIOUS SHIP GROUP

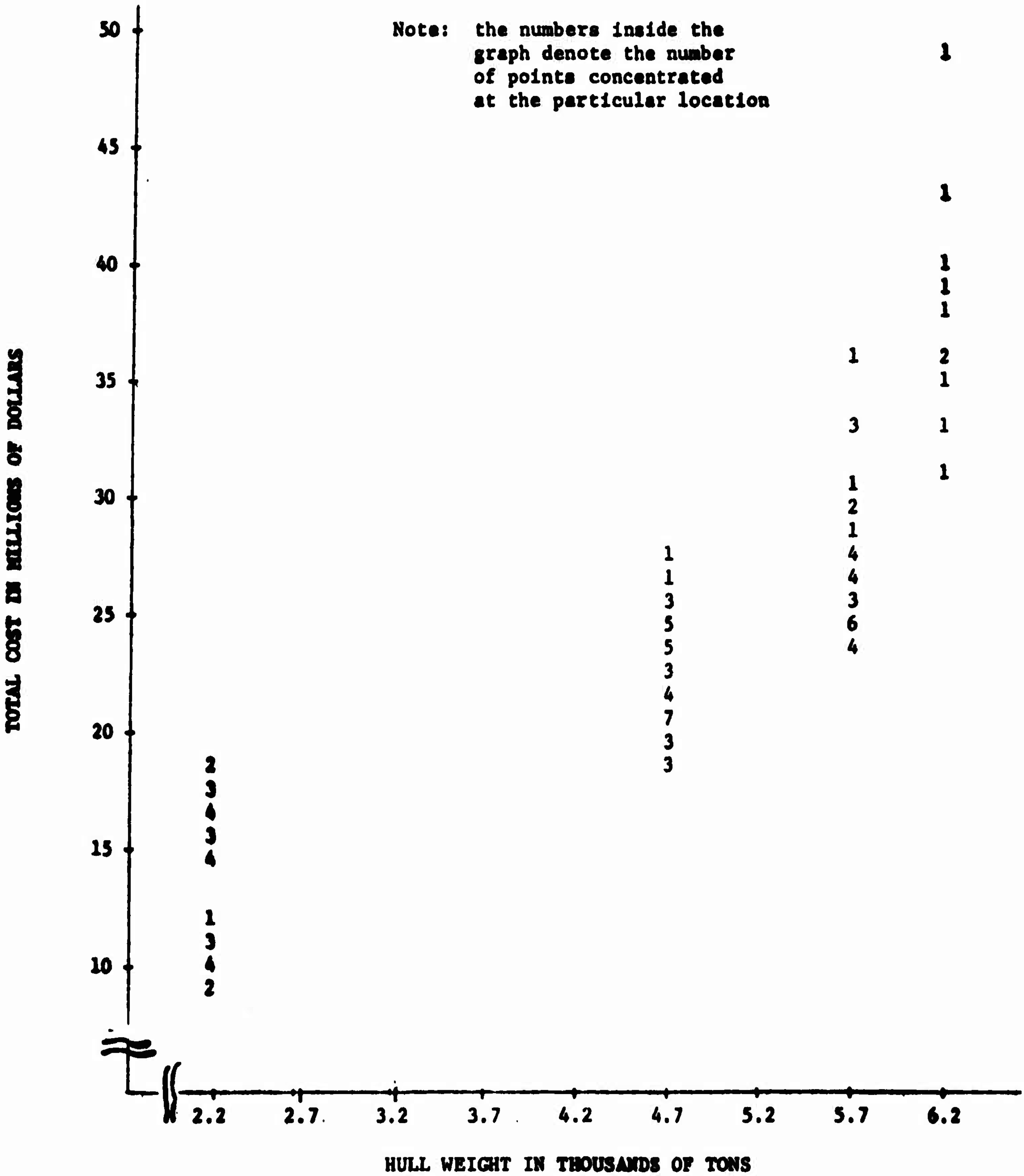
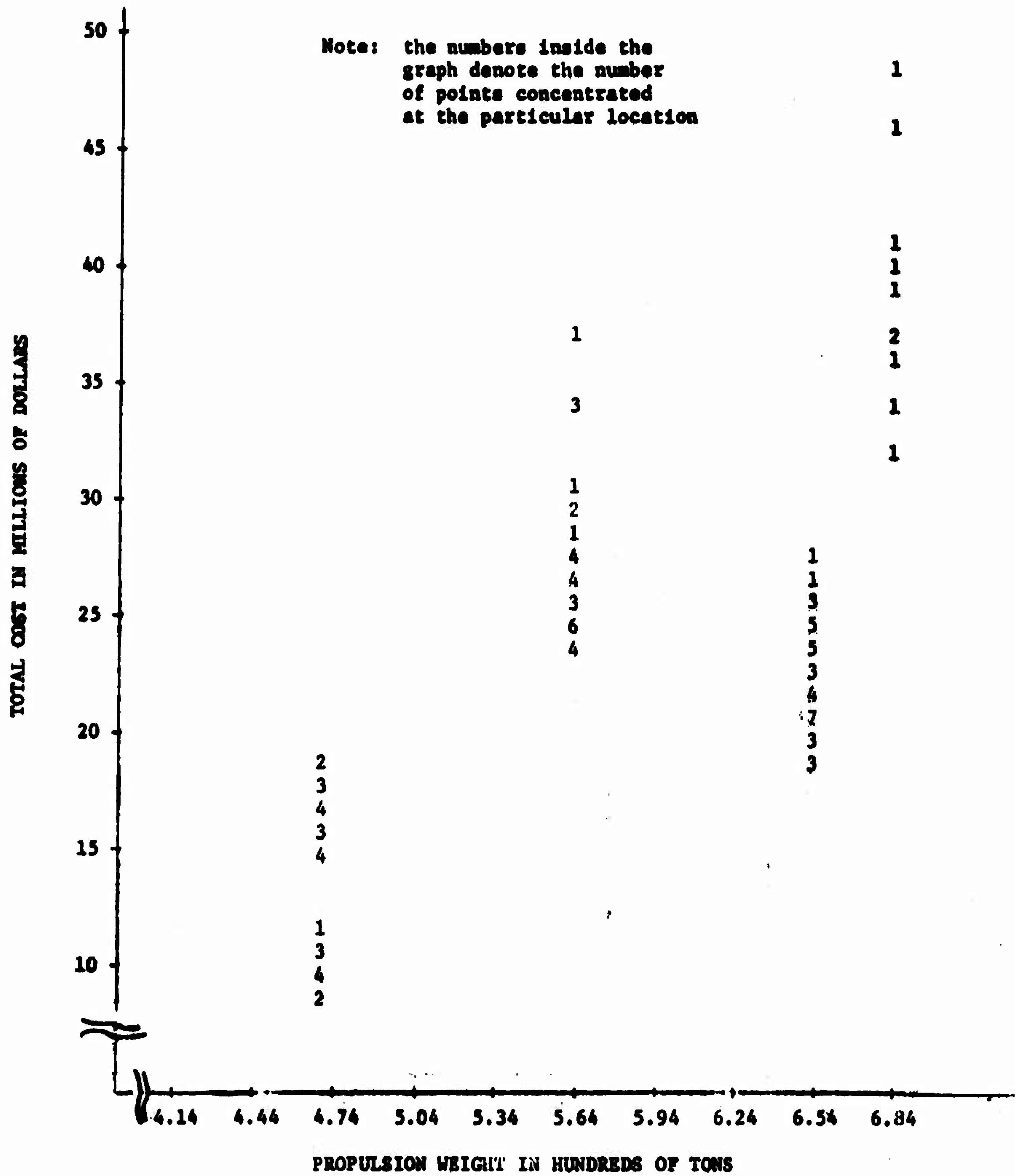


FIGURE 2

TOTAL COST VERSUS PROPULSION WEIGHT FOR THE AMPHIBIOUS SHIP GROUP



CHAPTER V

CONCLUSIONS AND RECOMMENDATIONS

The results of the various experiments in this study indicated that there are some significant relationships between the component weights of the ship and the total cost of the ship. Multiple regression equations which are statistically significant can be derived by using the total cost of the ship as the dependent variable and the component weights as the independent variables. However, the degree of accuracy with which these multiple regression equations predict the total cost of the ship leaves much to be desired. This degree of accuracy is highly dependent on how the samples used in the regression analysis were chosen. It was found that the chances of achieving a high degree of accuracy are best if the samples were selected in accordance with the following two guidelines: (1) The ships are grouped by usage, for example, amphibious ships, combat ships etc., (2) The sample standard deviation should be small relative to the sample mean. Following the above two guidelines, multiple regression equations yielding standard errors of estimate which are about 15% of the sample mean can be achieved. It should be noted that this degree of accuracy is not sufficient for practical prediction purposes. However, one has to take into consideration that there are two experimental limitations which exist in this study. First, the total cost of the ship which was used as the dependent variable was estimated by the contractors. It was not the true manufacturing cost. Second, the weights of the ship components which were used as

the independent variables were only approximate weight values. They were not the exact weight values. In addition to the above two limitations, another limitation is that due to lack of data. It was not possible in this study to perform experiments which further stratify ships into even more homogeneous groups, especially into samples consisting of ships belonging to only one ship type. It is suggested that if this stratification can be achieved, the standard error of estimate may be reduced significantly.

Numerous forms of non-linear prediction models have been tested, but none has shown any significant improvement over the linear models in its ability to predict the total cost of the ship.

One significant finding in this study is that for appropriately selected samples, where multiple regression equations which give comparatively better predictions can be derived, the adding-up process proves to be an inferior prediction tool for the total cost of the ship as compared to the multiple regression equation.

The overall results of this study indicated that statistical implementation of the engineering approach as represented by the regression analysis will not produce the kind of accuracy desired for ship cost estimation. But it does suggest that with the use of regression analysis, it is possible to produce a family of estimation functions or curves within which a certain range for the total cost of the ship can be established to enable the Navy to assess the credibility of a set of estimations made by contractors.

BIBLIOGRAPHY

Solomon, Henry. "Estimating cost of New Construction Navy Ships." Program in Logistics, George Washington University. April 8, 1969.

Solomon, Henry. "Methodological Problems in Estimating Costs of Shipbuilding Programs and Some Proposed Solutions." Program in Logistics, George Washington University, 1970.

Smillie, K. W. An Introduction to Regression and Correlation. New York: Academic Press, 1966.

Mendenhall, William. Introduction to Linear Models and Design and Analysis of Experiments. California: Wadsworth Publishing Company, Inc., 1968.

Johnston, J. Econometric Methods. New York: Mcgraw-Hill Book Company, Inc., 1963.

Dixaon, W. J. Biomedical Computer Programs. California: University of California Press.

Draper, Norman and Smith, Harry. Applied Regression Analysis. New York: John Wiley & Sons, Inc.

NONE
Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) THE GEORGE WASHINGTON UNIVERSITY Program in Logistics	2a. REPORT SECURITY CLASSIFICATION NONE 2b. GROUP

3. REPORT TITLE

A STATISTICAL ANALYSIS OF THE ENGINEERING APPROACH TO NAVY SHIPBUILDING COST ESTIMATION

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

SCIENTIFIC

5. AUTHOR(S) (Last name, first name, initial)

YU, K. C.

6. REPORT DATE 30 June 1970	7a. TOTAL NO. OF PAGES 67	7b. NO. OF REFS 7
8a. CONTRACT OR GRANT NO. N00014-67-A-0214 b. PROJECT NO. NR 347 020 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) T-240 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	

10. AVAILABILITY/LIMITATION NOTICES

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Office of Naval Research

13. ABSTRACT

In this study, the feasibility of developing regression models to predict the total cost of a Navy ship using the physical weights of the ship components as independent variables was investigated. The various forms of regression analyses fall under the following three categories:

(1) Linear multiple regression analysis.

(2) Non-line multiple regression analysis.

(3) Adding-up process, which is an aggregation of two-variable regression analyses.

It was found that the linear model is preferable over the non-linear model and the adding-up process. If the samples are properly selected, linear models which are statistically significant can be derived. Given its superiority over the other two models, the degree of accuracy of the linear model is still not high enough to produce a dependable point estimation for the total cost of the ship.

DD FORM 1 JAN 64 1473

NONE
Security Classification

NONE

Security Classification

14. KEY WORDS	LINK A ROLE	LINK A WT	LINK B ROLE	LINK B WT	LINK C ROLE	LINK C WT
Stepwise Regression Analysis						
Linear Regression Analysis						
Non-linear Regression Analysis						
Adding-up Process						
Standard Error of Estimate						
Coefficient of Variation						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through
____ ."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through
____ ."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through
____ ."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.